Generative Verifiers: Reward Modeling as Next-Token Prediction

Year: 2024 | Citations: 322 | Authors: Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar

Abstract

Verifiers or reward models are often used to enhance the reasoning performance of large language models (LLMs). A common approach is the Best-of-N method, where N candidate solutions generated by the LLM are ranked by a verifier, and the best one is selected. While LLM-based verifiers are typically trained as discriminative classifiers to score solutions, they do not utilize the text generation capabilities of pretrained LLMs. To overcome this limitation, we instead propose training verifiers using the ubiquitous next-token prediction objective, jointly on verification and solution generation. Compared to standard verifiers, such generative verifiers (GenRM) can benefit from several advantages of LLMs: they integrate seamlessly with instruction tuning, enable chain-of-thought reasoning, and can utilize additional test-time compute via majority voting for better verification. We demonstrate that GenRM outperforms discriminative, DPO verifiers, and LLM-as-a-Judge, resulting in large performance gains with Best-of-N, namely 5% $\rightarrow$ 45.3% on algorithmic tasks and 73% $\rightarrow$ 93.4% on GSM8K. In easy-to-hard generalization settings, we observe improvements of 28% $\rightarrow$ 44.6% on MATH, and 37.9% $\rightarrow$ 53.5% on MMLU abstract algebra. Furthermore, we find that training GenRM with synthetic verification rationales is sufficient to pick out subtle errors on math problems. Finally, we demonstrate that GenRM scales favorably with model size and test-time compute.